

ALGORITHMIC ROBUSTNESS FOR LEARNING VIA (ϵ, γ, τ) -GOOD SIMILARITY FUNCTIONS

Maria-Irina Nicolae, Marc Sebban & Amaury Habrard

Hubert Curien Laboratory

Jean Monnet University, Saint-Etienne, France

{Maria.Irina.Nicolae, Marc.Sebban, Amaury.Habrard}@univ-st-etienne.fr

Éric Gaussier & Massih-Reza Amini

Laboratoire d'Informatique de Grenoble

Joseph Fourier University, Grenoble, France

{Eric.Gaussier, Massih-Reza.Amini}@imag.fr

ABSTRACT

The notion of metric plays a key role in machine learning problems such as classification, clustering or ranking. However, it is worth noting that there is a severe lack of theoretical guarantees that can be expected on the generalization capacity of the classifier associated to a given metric. The theoretical framework of (ϵ, γ, τ) -good similarity functions (Balcan et al., 2008) has been one of the first attempts to draw a link between the properties of a similarity function and those of a linear classifier making use of it. In this paper, we extend and complete this theory by providing a new generalization bound for the associated classifier based on the algorithmic robustness framework.

1 INTRODUCTION

Most of the machine learning algorithms make use of metrics for comparing objects and making decisions (e.g. SVMs, k-NN, k-means, etc.). However, it is worth noticing that the theoretical guarantees of these algorithms are always derived independently from the peculiarities of the metric they make use of. For example, in supervised learning, the generalization bounds on the classification error do not take into account the discriminative properties of the metrics. In this context, Balcan et al. (2008) filled this gap by proposing the first framework that allows one to relate similarities with a classification algorithm. This general framework, that can be used with any bounded similarity function, provides generalization guarantees on a linear classifier learned from the similarity. Moreover, their algorithm, whose formulation is equivalent to a relaxed L_1 -norm SVM (Zhu et al., 2003), does not enforce the positive definiteness constraint of the similarity. In this paper, we show that using Balcan et al.'s setting and the algorithmic robustness framework (Xu & Mannor, 2012), we can derive generalization guarantees which consider other properties of the similarity. This leads to new consistency bounds for different kinds of similarity functions.

2 NOTATIONS AND RELATED WORK

Let us assume we are given access to labeled examples $\mathbf{z} = (\mathbf{x}, l(\mathbf{x}))$ drawn from some unknown distribution P over $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} = \{-1, 1\}$ are respectively the instance and the output spaces. A pairwise similarity function $K_{\mathbf{A}}$ over \mathcal{X} , possibly parameterized by a matrix $\mathbf{A} \in \mathbb{R}^{d \times d}$, is defined as $K_{\mathbf{A}} : \mathcal{X} \times \mathcal{X} \rightarrow [-1, 1]$, and the hinge loss as $[c]_+ = \max(0, 1 - c)$. We denote the L_1 norm by $\|\cdot\|_1$, the L_2 norm by $\|\cdot\|_2$ and the Frobenius norm by $\|\cdot\|_{\mathcal{F}}$. We assume that $\|\mathbf{x}\|_2 \leq 1$.

Balcan et al. (2008) introduced a theory for learning with so called (ϵ, γ, τ) -good similarity functions. Their generalization guarantees are based on the following definition.

Definition 1. (Balcan et al., 2008) $K_{\mathbf{A}}$ is a (ϵ, γ, τ) -good similarity function in hinge loss for a learning problem P if there exists a random indicator function $R(\mathbf{x})$ defining a probabilistic set of "reasonable points" such that the following conditions hold:

1. $\mathbb{E}_{(\mathbf{x}, l(\mathbf{x})) \sim P} [1 - l(\mathbf{x})g(\mathbf{x})/\gamma]_+ \leq \epsilon$,
where $g(\mathbf{x}) = \mathbb{E}_{(\mathbf{x}', l(\mathbf{x}'), R(\mathbf{x}'))} [l(\mathbf{x}')K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')|R(\mathbf{x}')]$.
2. $\Pr_{\mathbf{x}'}(R(\mathbf{x}')) \geq \tau$.

This definition imposes a constraint on the mass of reasonable points one must consider (greater than τ). It also expresses the tolerated margin violations in an averaged way: a $(1 - \epsilon)$ proportion of examples \mathbf{x} are on average 2γ more similar to random reasonable examples \mathbf{x}' of their own label than to random reasonable examples \mathbf{x}' of the other label. Definition 1 can then be used to learn well:

Theorem 1. (Balcan et al., 2008) Let $K_{\mathbf{A}}$ be an (ϵ, γ, τ) -good similarity function in hinge loss for a learning problem P . For any $\epsilon_1 > 0$ and $\delta < \gamma\epsilon_1/4$ let $\mathcal{S} = \{\mathbf{x}'_1, \mathbf{x}'_2, \dots, \mathbf{x}'_{d_u}\}$ be a sample of $d_u = \frac{2}{\tau} \left(\log(2/\delta) + 16 \frac{\log(2/\delta)}{(\epsilon_1\gamma)^2} \right)$ landmarks drawn from P . Consider the mapping $\phi^{\mathcal{S}} : \mathcal{X} \rightarrow \mathbb{R}^{d_u}$ defined as follows: $\phi^{\mathcal{S}}_i(\mathbf{x}) = K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}'_i), i \in \{1, \dots, d_u\}$. With probability $1 - \delta$ over the random sample \mathcal{S} , the induced distribution $\phi^{\mathcal{S}}(P)$ in \mathbb{R}^{d_u} , has a separator achieving hinge loss at most $\epsilon + \epsilon_1$ at margin γ .

In other words, if $K_{\mathbf{A}}$ is (ϵ, γ, τ) -good according to Definition 1 and enough points are available, there exists a linear separator $\alpha \in \mathbb{R}^{d_u}$ with error arbitrarily close to ϵ in the space $\phi^{\mathcal{S}}$. This separator can be learned from d_l labeled examples by solving the following optimization problem:

$$\min \frac{1}{d_l} \sum_{i=1}^{d_l} \ell(\mathbf{A}, \alpha, \mathbf{z}_i) \quad \text{s.t.} \quad \sum_{j=1}^{d_u} |\alpha_j| \leq 1/\gamma \quad (1)$$

where $\ell(\mathbf{A}, \alpha, \mathbf{z}_i = (\mathbf{x}_i, l(\mathbf{x}_i))) = \left[1 - \sum_{j=1}^{d_u} \alpha_j l(\mathbf{x}_i) K_{\mathbf{A}}(\mathbf{x}_i, \mathbf{x}'_j) \right]_+$ is the instantaneous loss estimated at point $(\mathbf{x}_i, l(\mathbf{x}_i))$. Therefore, this optimization problem reduces to minimizing the empirical loss $\hat{R}^\ell = \frac{1}{d_l} \sum_{i=1}^{d_l} \ell(\mathbf{A}, \alpha, \mathbf{z}_i)$ over the training set \mathcal{S} . Note that this problem can be solved efficiently by linear programming. Also, as the problem is L_1 -constrained, tuning the value of γ will produce a sparse solution.

3 CONSISTENCY GUARANTEES

In this section, we provide a new generalization bound for the classifier learned in Problem (1) based on the recent algorithmic robustness framework proposed by Xu & Mannor (2012). To begin with, let us recall the notion of robustness of an algorithm \mathcal{A} .

Definition 2 (Algorithmic Robustness (Xu & Mannor, 2012)). Algorithm \mathcal{A} is $(M, \epsilon(\cdot))$ -robust, for $M \in \mathbb{N}$ and $\epsilon(\cdot) : \mathcal{Z}^{d_l} \rightarrow \mathbb{R}$, if \mathcal{Z} can be partitioned into M disjoint sets, denoted by $\{C_i\}_{i=1}^M$, such that the following holds for all $\mathcal{S} \in \mathcal{Z}^{d_l}$:

$$\begin{aligned} \forall \mathbf{z} = (\mathbf{x}, l(\mathbf{x})) \in \mathcal{S}, \forall \mathbf{z}' = (\mathbf{x}', l(\mathbf{x}')) \in \mathcal{Z}, \forall i \in [M] : \\ \text{if } \mathbf{z}, \mathbf{z}' \in C_i, \text{ then } |\ell(\mathbf{A}, \alpha, \mathbf{z}) - \ell(\mathbf{A}, \alpha, \mathbf{z}')| \leq \epsilon(\mathcal{S}). \end{aligned}$$

Roughly speaking, robustness characterizes the capability of an algorithm to perform similarly on close train and test instances. The closeness of the instances is based on a partitioning of \mathcal{Z} : two examples are close if they belong to the same region. In general, the partition is based on the notion of covering number (Kolmogorov & Tikhomirov, 1961) allowing one to cover \mathcal{Z} by regions where the distance/norm between two elements in the same region are no more than a fixed quantity ρ (see Xu & Mannor (2012) for details about how the covering is built). Now we can state the first theoretical contribution of this paper.

Theorem 2. Given a partition of \mathcal{Z} into M subsets $\{C_i\}$ such that $\mathbf{z} = (\mathbf{x}, l(\mathbf{x}))$ and $\mathbf{z}' = (\mathbf{x}', l(\mathbf{x}')) \in C_i$ and $l(\mathbf{x}) = l(\mathbf{x}')$, and provided that $K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}')$ is l -lipschitz w.r.t. its first argument, the optimization problem (1) is $(M, \epsilon(\mathcal{S}))$ -robust with $\epsilon(\mathcal{S}) = \frac{1}{\gamma}l\rho$, where $\rho = \sup_{\mathbf{x}, \mathbf{x}' \in C_i} \|\mathbf{x} - \mathbf{x}'\|$.

Proof.

$$|\ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}) - \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}')| \leq \left| \sum_{j=1}^{d_u} \alpha_j l(\mathbf{x}') K_{\mathbf{A}}(\mathbf{x}', \mathbf{x}_j) - \sum_{j=1}^{d_u} \alpha_j l(\mathbf{x}) K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}_j) \right| \quad (2)$$

$$\leq \sum_{j=1}^{d_u} |\alpha_j| \cdot |K_{\mathbf{A}}(\mathbf{x}', \mathbf{x}_j) - K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}_j)| \quad (3)$$

$$\leq \sum_{j=1}^{d_u} |\alpha_j| \cdot l \|\mathbf{x} - \mathbf{x}'\| \leq \frac{1}{\gamma} l \rho \quad (4)$$

Setting $\rho = \sup_{\mathbf{x}, \mathbf{x}' \in C_i} \|\mathbf{x} - \mathbf{x}'\|_1$, we get the Theorem. We get Inequality (2) from the 1-lipschitzness of the hinge loss; Inequality (3) comes from the classical triangle inequality; The first inequality on line (4) is due to the l -lipschitzness of $K_{\mathbf{A}}(\mathbf{x}, \mathbf{x}_j)$ and the result follows from the constraint of Problem (1). \square

We now give a PAC generalization bound on the true loss making use of the previous robustness result. Let $\mathcal{R}^\ell = \mathbb{E}_{\mathbf{z} \sim \mathcal{Z}} \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z})$ be the true loss w.r.t. the unknown distribution \mathcal{Z} and $\hat{\mathcal{R}}^\ell = \frac{1}{d_l} \sum_{i=1}^{d_l} \ell(\mathbf{A}, \boldsymbol{\alpha}, \mathbf{z}_i)$ be the empirical loss over the training set \mathcal{S} .

Theorem 3. Considering that problem (1) is $(M, \epsilon(\mathcal{S}))$ -robust, and that $K_{\mathbf{A}}$ is l -lipschitz w.r.t. to its first argument, for any $\delta > 0$ with probability at least $1 - \delta$, we have:

$$|\mathcal{R}^\ell - \hat{\mathcal{R}}^\ell| \leq \frac{1}{\gamma} l \rho + B \sqrt{\frac{2M \ln 2 + 2 \ln(1/\delta)}{d_l}},$$

where $B = 1 + \frac{1}{\gamma}$ is an upper bound of the loss ℓ .

The proof of Theorem 3 follows the one described in Xu & Mannor (2012) and makes use of a concentration inequality over multinomial random variables (van der Vaart & Wellner, 1996). Note that in robustness bounds, the cover radius ρ can be made arbitrarily small at the expense of larger values of M . As M appears in the second term, which decreases to 0 when d_l tends to infinity, this bound provides a standard $O(1/\sqrt{d_l})$ asymptotic convergence.

The previous theorem strongly depends on the l -lipschitzness of the similarity function. In the following, we focus on some particular similarities that can be used in this setting: $K_{\mathbf{A}}^1$, a similarity derived from the Mahalanobis distance, $K_{\mathbf{A}}^2$ a bilinear similarity and $K_{\mathbf{A}}^3$ an exponential similarity. We provide the proof of the l -lipschitzness for $K_{\mathbf{A}}^1$. The two others follow the same ideas.

Similarity function 1. We define $K_{\mathbf{A}}^1(\mathbf{x}, \mathbf{x}') = 1 - (\mathbf{x} - \mathbf{x}')^T \mathbf{A}(\mathbf{x} - \mathbf{x}')$, a similarity derived from the Mahalanobis distance. $K_{\mathbf{A}}^1(\mathbf{x}, \mathbf{x}')$ is $4\|\mathbf{A}\|_2$ -lipschitz w.r.t. its first argument.

Proof.

$$\begin{aligned} |K_{\mathbf{A}}^1(\mathbf{x}, \mathbf{x}'') - K_{\mathbf{A}}^1(\mathbf{x}', \mathbf{x}'')| &= |1 - ((\mathbf{x} - \mathbf{x}'')^T \mathbf{A}(\mathbf{x} - \mathbf{x}'')) - 1 + ((\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x}' - \mathbf{x}''))| \\ &= |(\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x}' - \mathbf{x}'') - (\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x} - \mathbf{x}'') + (\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x} - \mathbf{x}'') - (\mathbf{x} - \mathbf{x}'')^T \mathbf{A}(\mathbf{x} - \mathbf{x}'')| \\ &= |(\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x}' - \mathbf{x}) + (\mathbf{x}' - \mathbf{x})^T \mathbf{A}(\mathbf{x} - \mathbf{x}'')| \\ &\leq |(\mathbf{x}' - \mathbf{x}'')^T \mathbf{A}(\mathbf{x}' - \mathbf{x})| + |(\mathbf{x}' - \mathbf{x})^T \mathbf{A}(\mathbf{x} - \mathbf{x}'')| \\ &\leq \|\mathbf{x}' - \mathbf{x}''\|_2 \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{x}' - \mathbf{x}\|_2 + \|\mathbf{x}' - \mathbf{x}\|_2 \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{x} - \mathbf{x}''\|_2 \quad (5) \\ &\leq \|\mathbf{x}' - \mathbf{x}''\|_2 \cdot \|\mathbf{A}\|_2 \cdot (\|\mathbf{x}'\|_2 + \|\mathbf{x}\|_2) + \|\mathbf{x}' - \mathbf{x}\|_2 \cdot \|\mathbf{A}\|_2 \cdot (\|\mathbf{x}\|_2 + \|\mathbf{x}''\|_2) \\ &\leq 4 \cdot \|\mathbf{A}\|_2 \cdot \|\mathbf{x} - \mathbf{x}'\|. \quad (6) \end{aligned}$$

Inequality (5) comes from the Cauchy-Schwarz inequality and some classical norm properties; Inequality (6) comes from the assumption that $\|\mathbf{x}\|_2 \leq 1$. \square

Similarity function 2. Let $K_{\mathbf{A}}^2$ be the bilinear form $K_{\mathbf{A}}^2(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{A} \mathbf{x}'$. $K_{\mathbf{A}}^2(\mathbf{x}, \mathbf{x}')$ is $\|\mathbf{A}\|_2$ -lipschitz w.r.t. its first argument.

Similarity function 3. Let $K_{\mathbf{A}}^3(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{(\mathbf{x}-\mathbf{x}')^T \mathbf{A} (\mathbf{x}-\mathbf{x}')}{2\sigma^2}\right)$. $K_{\mathbf{A}}^3(\mathbf{x}, \mathbf{x}')$ is l -lipschitz w.r.t. its first argument with $l = \frac{2\|\mathbf{A}\|_2}{\sigma^2} \left(\exp\left(\frac{1}{2\sigma^2}\right) - \exp\left(\frac{-1}{2\sigma^2}\right)\right)$.

As both $K_{\mathbf{A}}^1$ and $K_{\mathbf{A}}^2$ are linear w.r.t. their arguments, they have the main advantage to keep problem (1) convex. $K_{\mathbf{A}}^3$ is also based on the Mahalanobis distance, but this time it is a non linear function, resembling more a gaussian kernel. Plugging $l = 4\|\mathbf{A}\|_2$ (resp. $l = \|\mathbf{A}\|_2$ and $l = \frac{2\|\mathbf{A}\|_2}{\sigma^2} \left(\exp\left(\frac{1}{2\sigma^2}\right) - \exp\left(\frac{-1}{2\sigma^2}\right)\right)$) in Theorem 3, we obtain consistency results for problem (1) using $K_{\mathbf{A}}^1(\mathbf{x}, \mathbf{x}')$ (resp. $K_{\mathbf{A}}^2(\mathbf{x}, \mathbf{x}')$ and $K_{\mathbf{A}}^3(\mathbf{x}, \mathbf{x}')$). As the gap between empirical and true loss presented in Theorem 3 is proportional with l for the l -lipschitzness of each similarity function, we would like to keep this parameter as small as possible. We notice that the generalization bound is tighter for $K_{\mathbf{A}}^1$ than for $K_{\mathbf{A}}^2$. The bound for $K_{\mathbf{A}}^3$ depends on the additional parameter σ , that adjusts the influence of the similarity value w.r.t. the distance to the landmarks. The value of l goes to 0 as σ augments, so larger values of σ are preferable in order to obtain a tight bound for the generalization error. However, note that when σ is large, the exponential behaves almost linearly, i.e. the projection loses its non-linear power.

4 CONCLUSION

In this paper, we extended the theoretical analysis of the (ϵ, γ, τ) -good similarity framework. Using the algorithmic robustness setting, we derived new generalization bounds for different similarity functions. It turns out that the smaller the lipschitz constant of those similarity functions, the tighter the consistency bounds. This opens the door to new lines of research in *metric learning* (Bellet et al., 2013; 2015) aiming at maximizing the (ϵ, γ, τ) -goodness of similarity functions s.t. $\|\mathbf{A}\|_2$ is as small as possible (see pioneer works like Bellet et al. (2012; 2011)).

ACKNOWLEDGEMENTS

Funding for this project was provided by a grant from Région Rhône-Alpes.

REFERENCES

- Balcan, M.-F., Blum, A., and Srebro, N. Improved guarantees for learning via similarity functions. In Servedio, R. A. and Zhang, T. (eds.), *COLT*, pp. 287–298. Omnipress, 2008.
- Bellet, A., Habrard, A., and Sebban, M. Learning good edit similarities with generalization guarantees. In *ECML PKDD 2011*, pp. 188–203, 2011.
- Bellet, A., Habrard, A., and Sebban, M. Similarity learning for provably accurate sparse linear classification. In *ICML 2012*, pp. 1871–1878, 2012.
- Bellet, A., Habrard, A., and Sebban, M. A survey on metric learning for feature vectors and structured data. *arXiv preprint arXiv:1306.6709*, 2013.
- Bellet, A., Habrard, A., and Sebban, M. *Metric Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers, 2015.
- Kolmogorov, A. and Tikhomirov, V. ϵ -entropy and ϵ -capacity of sets in functional spaces. *American Mathematical Society Translations*, 2(17):277–364, 1961.
- van der Vaart, A. and Wellner, J. *Weak Convergence and Empirical Processes*. Springer series in statistics. Springer, 1996.
- Xu, H. and Mannor, S. Robustness and generalization. *Machine Learning*, 86(3):391–423, 2012.
- Zhu, J., Rosset, S., Hastie, T., and Tibshirani, R. 1-norm support vector machines. In *NIPS*, pp. 16. MIT Press, 2003.